Hindawi
Journal of Food Quality
Volume 2018, Article ID 3136516, 11 pages
https://doi.org/10.1155/2018/3136516

Research Article

Key Variables Screening of Near-Infrared Models for Simultaneous Determination of Quality Parameters in Traditional Chinese Food "Fuzhu"

Jiahua Wang,[1] **Jun Wang**,[2] **Xiaowei Zhang**,[2] **Jingjing Cheng**,[2] **and Qingyu Li**[2]

[1]*College of Food Science and Engineering, Wuhan Polytechnic University, Wuhan, Hubei 430023, China*
[2]*College of Food & Bioengineering, Xuchang University, Xuchang, Henan 461000, China*

Correspondence should be addressed to Jiahua Wang; w.jiahua@163.com

Received 8 September 2017; Accepted 8 February 2018; Published 15 March 2018

Academic Editor: Elena González-Fandos

The traditional Chinese food Fuzhu is a dried soy protein-lipid film formed during the heating of soymilk. This study investigates whether a simple and accurate model can nondestructively determine the quality parameters of intact Fuzhu. The diffused reflectance spectra (1000–2499 nm) of intact Fuzhu were collected by a commercial near-infrared (NIR) spectrometer. Among various preprocessing methods, the derivative by wavelet transform method optimally enhanced the characteristic signals of Fuzhu spectra. Uninformative variable elimination based on Monte Carlo (MC-UVE), random frog (RF), and competitive adaptive reweighted sampling (CARS) were proposed to select key variables for partial least squares (PLS) calculation. The strong performance of the developed models is attributed to the high ratios of prediction to deviation values (3.32–3.51 for protein, 3.62–3.89 for lipid, and 4.27–4.55 for moisture). The prediction set was used to assess the performances of the best models of protein (CARS-PLS), lipid (RF-PLS), and moisture (CARS-PLS), which resulted in greater coefficients of determination of 0.958, 0.966, and 0.976, respectively, and lower root mean square errors of prediction of 0.656%, 0.442%, and 0.123%, respectively. Combined with chemometrics methods, the NIR technique is promising for simultaneous testing of quality parameters of intact Fuzhu.

1. Introduction

The Chinese traditional soybean food "Fuzhu" originates from the Tang Dynasty in ancient China and has long been considered a luxury in China and Japan. Formally described as dried soybean protein-lipid film, Fuzhu is formed during the heating of soymilk [1, 2]. Fuzhu is increasingly sought by domestic and overseas consumers because of its meatlike texture and high nutritional value. Along with the increased consumption, large-scale production of this food has gradually emerged in recent years. The yield per year has reached 200,000 tons in China. To protect and ensure the inheritance of traditional Chinese food, the Administration of Quality Supervision, Inspection and Quarantine (AQSIQ) of China has authenticated the Fuzhu produced in Xuchang, Gaoan, Tuodong, and Guilin (China's main production areas) as National Geographical Indication Protection Product (NGIPP). The AQSIQ has imposed standards on the quality parameters (protein, lipid, and moisture contents) of NGIPPs. The protein, lipid, and moisture contents provide valuable information for commercial pricing, because they underlie the required quality standards of Fuzhu foods sold to consumers. However, current chemical analysis methods are time-consuming and costly, requiring the extensive use of auxiliary chemicals. Therefore, a rapid, high-efficiency, online analytical method is urgently required.

Combined with chemometrics, near-infrared (NIR) spectroscopy is a fast, accurate, nondestructive technique that is easily implemented. Moreover, a series of professional reviews have reported the wide usage of the NIR technique in the analysis of quality parameters in food industries [3–6]. In practical applications, the NIR spectra always derive from meaningful variables contributed by sample attributes and from noise variables caused by environmental and instrumental fluctuations. Therefore, the spectral data must be pretreated to improve the important signal characteristics.

Commonly used preprocessing methods are Savitzky Golay (SG) smoothing [7], the Norris derivative filter (NDF), first-order derivative (1D) or second-order derivative (2D), multiplicative scatter correction (MSC), and the standard normal variate (SNV) [8]. These methods, respectively, filter the high-frequency noises, improve the signal-to-noise ratio, enhance the spectroscopy resolution, resolve the overlapping peaks, correct the baseline, and eliminate scatter. As an alternative to these conventional methods, Shao and Ma [9] and Nie et al. [10] applied the wavelet transform (WT) to the derivative calculation (D-WT). Singh et al. [11] reported the theory of WT analysis and reviewed its application in signal processing and feature extraction for quality monitoring of agricultural and food products.

Variable selection is a critical step when analyzing datasets with hundreds of thousands of variables in NIR spectroscopy. Thus, selecting the key variables is essential for improving the efficiency and decreasing the prediction errors in a robust model [12–15]. Variable selection also simplifies the model dimensionality, improves the interpretation, and lowers the measurement costs [16]. The performance of partial least squares (PLS) models has been enhanced by numerous variable selection methods, such as genetic algorithms (GAs), uninformative variable elimination based on Monte Carlo (MC-UVE), random frog (RF), successive projections algorithms (SPAs), variable importance in projection (VIP), and competitive adaptive reweighted sampling (CARS). Variable selection by MC-UVE combined with GAs has been applied in predictions of soluble solid contents in watermelon [17] and the prior storage period of lamb's lettuce [18]. The RF proved to be a promising selector of cancer-related genes [19]. The RF algorithm was also proposed as the variable selector in the determination of polyphenol contents of tea from 14 tea-tree cultivars [20] and the detection of fungus infection on rapeseed petals [21]. Zhang et al. [22] predicted the pH of anaerobic digestion liquid of water hyacinth-rice straw mixtures by hyperspectral imaging. For this purpose, they selected 8, 15, and 20 optimal wavelengths using the SPAs, RF, and VIP, respectively. CARS was proposed for variable selection in calibration models of branched-amino acid contents (leucine, isoleucine, and valine) in fermented mycelia of the Chinese caterpillar fungus *(Cordyceps sinensis)* [23] and caffeine contents in roasted Arabica coffee [24]. However, to our knowledge, key variables selection by NIR modeling for simultaneous determination of the quality parameters in Fuzhu has never been reported.

The present work applies NIR spectroscopy to the simultaneous determination of protein, lipid, and moisture contents in intact Fuzhu. The specific objectives were (1) to determine a suitable pretreatment method for spectral processing; (2) to select the key variables for protein, lipid, and moisture analyses by MC-UVE, FR, and CARS; (3) to develop PLS models and inspect their practical performances.

2. Materials and Methods

2.1. Fuzhu Sample. Fuzhu samples were collected from local market in Xuchang, which is one of the China's main production areas. To create a wide range of protein, lipid, and moisture contents, Fuzhu samples with different prices or brands were deliberately selected. In total, 180 Fuzhu samples (about 500 g of each sample) were obtained from October 2015 to September 2016 and immediately transported to a cold storage contained at ~4°C with relative humidity of ~50% until the trial began.

2.2. Diffuse Reflectance Spectra Acquisition. Prior to NIR spectra acquisition, the Fuzhu samples were kept in laboratory (~24°C, ~62% relative humidity) for more than 8 h for temperature equilibration and to diminish the influence of temperature on the NIR spectral profile. After reaching equilibrium, the diffused reflectance NIR spectra ($\log(1/R)$) of intact Fuzhu were acquired by a commercially available NIR spectrometer (mode SupNIR-2750; FPI-INC Co., Hangzhou, China) equipped with an InGaAs detector. To acquire representative spectral information, a rotator with a motor-driven sample cup rotated during spectrum collection. Each spectrum was the average of 16 scanned spectra. $\log(1/R)$ spectra were recorded using the NIR spectrometer at 1 nm intervals, and the NIR wavelength ranged from 1000 to 2499 nm (giving in 1500 variables). For each spectrum determination, about 70 g Fuzhu samples were manually placed on the sample cup. Seven spectra were obtained for each Fuzhu sample by nonrepetitively loading the sample seven times. These spectra were averaged to obtain the final spectrum of each sample.

2.3. Reference Values Measurement. Before reference measurement, intact Fuzhu samples were crushed into powder using a grinder for consistent measurements. The powder particle size was kept below 40 Taylor mesh, and the sieved powders were collected for wet chemical analysis. The reference values of Fuzhu quality parameters were measured according to the national standards of China. The total protein content was determined by the classical Kjeldahl method using a Digestion Unit (mode DT 208; Foss Scino, Denmark) combined with a Kjeltec Analyzer Unit (mode 2300; Foss Tecator, Sweden), according to GB 5009.5-2010 [25]. The lipid content was measured by a Soxhlet Analyzer (mode SOX406; Hanon Instrument Co., Jinan, China) according to GB/T 5009.6-2003 [26]. The moisture content was determined in a drying oven (mode DZF-6020; Zhongxingweiye Co., Beijing, China), following GB 5009.3-2010 [27].

2.4. Chemometrics

2.4.1. Spectral Preprocessing Methods. Before model development, the original spectral data were subjected to spectral preprocessing using MSC, SNV, SG smoothing, NDF, 1D or 2D, and D-WT. In present study, the data point and the polynomial order of SG smoothing were set to 7 and 3, respectively. The segment length and the gap between segments of NDF were set as 5 and 5. Haar wavelet function with scale of 60 was used for D-WT calculating of analytical signals. The lowest root mean squared error of cross-validation (RMSECV) was used to determine suitable preprocessing method.

2.4.2. MC-UVE. MC-UVE, which combines uninformative variable elimination (UVE) with the Monte Carlo (MC) method, was proposed by Cai et al. [28] and is used for variable selection in NIR spectral modeling. The main steps of the MC-UVE procedure are described below:

(1) Construct an original sample set comprising an $(n \times p)$ spectral matrix X and an $(n \times 1)$ reference value matrix Y.

(2) Based on the MC technique, randomly select n_t samples from the original sample set. These samples constitute the training subset for building a PLS submodel. Record the regression coefficient (RC) of each variable. Specifically, if $\beta(\beta_1, \beta_2, \ldots, \beta_p)$ is the regression vector of the PLS model, the RCs are the $\beta_1, \beta_2, \ldots, \beta_p$ of the 1st, 2nd, $\ldots$, pth variable, respectively.

(3) Repeat Step (2) N times, and construct an $(N \times p)$ matrix β of the PLS RCs. Finally, calculate the reliability index (RI) of each variable by

$$\mathrm{RI} = \frac{\mathrm{mean}\left(\beta_j\right)}{\mathrm{std}\left(\beta_j\right)}, \quad j = 1, 2, 3, \ldots, p, \tag{1}$$

where $\mathrm{mean}(\beta_j)$ and $\mathrm{std}(\beta_j)$ denote the mean and standard deviation of the RC of the jth variable, respectively.

2.4.3. RF. RF is a simple and efficient method that borrows the framework of reversible-jump Markov chain Monte Carlo method. PLS is the modeling method in the RF procedure. The principles of RF are detailed in Li et al. [19] and Yun et al. [29]. Here, we summarize the main steps of the RF procedure:

(1) Construct an original sample set consisting of X and Y.

(2) Set the tuning parameters that control the RF performance and initialize a subset V_0 consisting of Q randomly selected variables.

(3) Generate a random number Q^* from the presupposed normal distribution $(Q, \theta Q)$, where θ is a constant. A candidate subset V^* of Q^* variables can be proposed in three situations, as detailed in the literature [29]. Retain the Q^* variables with the largest absolute RCs in the PLS model and collect them into the candidate subset $\mathbf{V}^*$.

(4) Using subsets V_0 and V^* in the PLS model, calculate the root mean squared error of cross-validation (RMSECV) and RMSECV*, respectively. If RMSECV$^* \leq$ RMSECV, accept V^* as V_1; otherwise, accept V^* as V_1 with a probability of $\eta\,\mathrm{RMSECV}/\mathrm{RMSECV}^*$, where η is a constant. Repeat Steps (3)–(5) N times, updating the candidate subset at each iteration.

(5) After N simulations, N subsets are obtained. Let N_j be the frequency at which the jth variable is selected in these N subsets. The selection probability (SP) of the jth variable is then computed by

$$\mathrm{SP}_j = \frac{N_j}{N}, \quad j = 1, 2, 3, \ldots, p. \tag{2}$$

2.4.4. CARS. Developed by Li et al. [30], CARS is a simple and effective method that selects the optimal combination of key variables of multicomponent spectral data. CARS is based on the "survival of the fittest" principle in Darwin's Theory of Evolution. The main steps of the CARS procedure are summarized below:

(1) Construct an original sample set consisting of X and Y.

(2) Using the MC technique, select n_0 samples (usually 80–90% of the original sample set), and construct them into an initialized subset V_0.

(3) Develop a PLS submodel based on V_0, and weight each variable by recording its RC. Define the importance of each wavelength variable by assigning a normalized weight as follows:

$$w_j = \frac{b_j}{\mathrm{sum}\left(b_j\right)}, \quad j = 1, 2, 3, \ldots, p, \tag{3}$$

where w_i denotes the normalized weight of the jth variable, and b_j is the absolute RC of the jth variable.

(4) Remove the wavelengths with small absolute RC by applying an exponentially decreasing function (EDF). The number of retained variables is computed by

$$m_i = p \cdot r_i = p \cdot a e^{-ki}, \quad i = 1, 2, 3, \ldots, N, \tag{4}$$

where a and k are two constants determined by two equations in Li et al. [30], and p represents the total number of variables.

(5) After reducing the number of wavelengths by EDF, eliminate weakly competitive variables (those with low weights) by adaptive reweighted sampling (ARS). The variables with dominant weights are retained for the PLS model construction. The new set V_1 containing m_i variables is considered as V_0 in Step (2) of the new CARS operation. Steps (2)–(5) are looped N times, updating the initialized subset at each iteration.

2.4.5. Operating Parameters and Process. Figure 1 is a flowchart of the MC-UVE, RF, and CARS algorithms. All maximum factors of PLS (A) were set to 10, and the data processing method in the three procedures was decided as "center." The ratio of training to total samples in the MC-UVE procedure was set to 0.75 (i.e., $n_t = 0.75n$). The number of MC simulations (N) was set to 500, sufficient for precise stability estimates. The suitable number of principal factors was determined by Monte Carlo cross-validation (MCCV) with F test methods. In this work, the mean RI values of the variables were obtained from 100 MC-UVE iterations. The mean RIs of all variables were ranked from highest to lowest. Variables with RIs below a certain threshold were eliminated.

In the RF procedure, the variables θ, ω, and η were set to their default values (0.3, 3, and 0.1, resp.) [19]. The number of simulations (N) and initialized variables (Q) were preset to 10000 and 20, respectively. Here, the mean SP values of each variable were obtained from 100 independent operations of the RF procedure. Again, the mean SPs of all variables were ranked from highest to lowest, and the SP of the n_jth variable was assigned as the threshold value. Variables with SP values above the threshold were retained for the PLS model calibration.

In the present work, the number of sampling runs before running the CARS procedure was set to 100. The RMSECVs

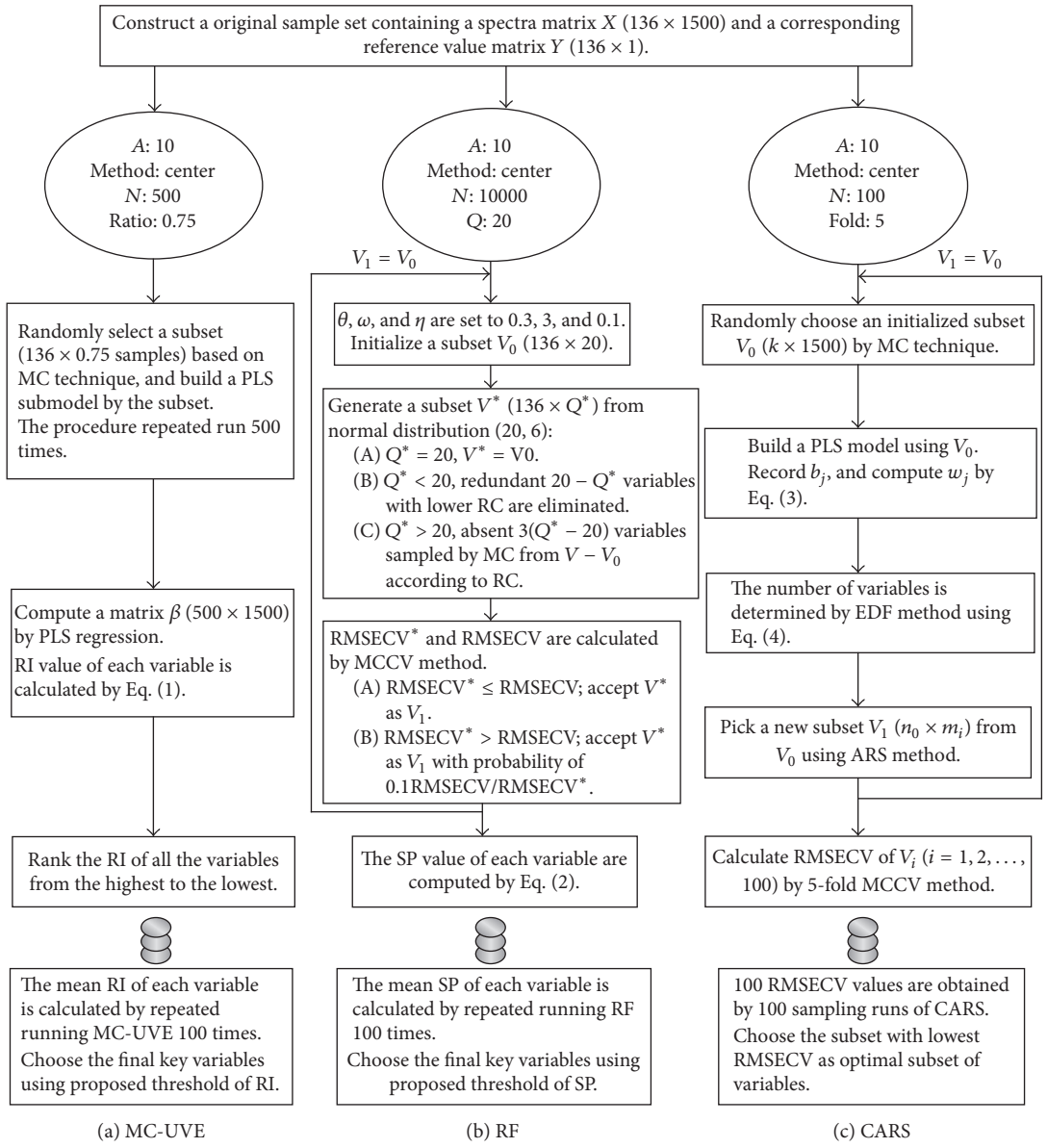

Figure 1: Flowchart of the MC-UVE, RF, and CARS algorithms.

of the 100 PLS models obtained by the 100 sampling runs of CARS were computed by the 5-fold MCCV method. Finally, the subset with lowest RMSECV was chosen as the optimal subset of variables.

The MC-UVE, RF, and CARS algorithms are contained in the libpls_1.95 toolbox, which is freely downloadable from http://www.libpls.net/download.php. All preprocessing procedures and the MC-UVE, RF, and CARS procedures were performed in MATLAB 7.10.0 (R2010a) (Math Works Inc., Natick, MA, USA).

2.5. Statistical Analysis. The resulting calibration equations between the chemical analyses and the NIR spectroscopy were evaluated by the coefficient of determination for calibration (R_c) and the root mean square error of calibration (RMSEC). The predictive precision was evaluated by the coefficient of determination for prediction (R_P) and the root mean square error of prediction (RMSEP). The RPD, which defines the ratio of the standard deviation (SD) in the prediction set to the RMSEP, has been previously used in model evaluation [31]. The coefficients of determination, RMSE indices, and RPD are, respectively, calculated by equations reported by Wang et al. [32].

3. Results and Discussion

3.1. Definitions of the Calibration and Prediction Sets. Based on the reference value, the 180 samples were split by the ranking method into two groups, the calibration and prediction sets (at an approximate ratio of 3 : 1). The calibration set (136

Table 1: Statistical results of calibration and prediction sets of Fuzhu quality parameters.

Quality indices	Calibration set				Prediction set			
	Number	Range (%)	Mean (%)	SD	Number	Range (%)	Mean (%)	SD
Protein	136	40.46–50.15	46.21	2.36	44	40.92–49.49	46.18	2.30
Lipid	136	16.95–25.85	20.90	1.86	44	17.81–25.07	20.84	1.72
Moisture	136	6.43–9.39	7.81	0.61	44	6.70–9.05	7.79	0.56

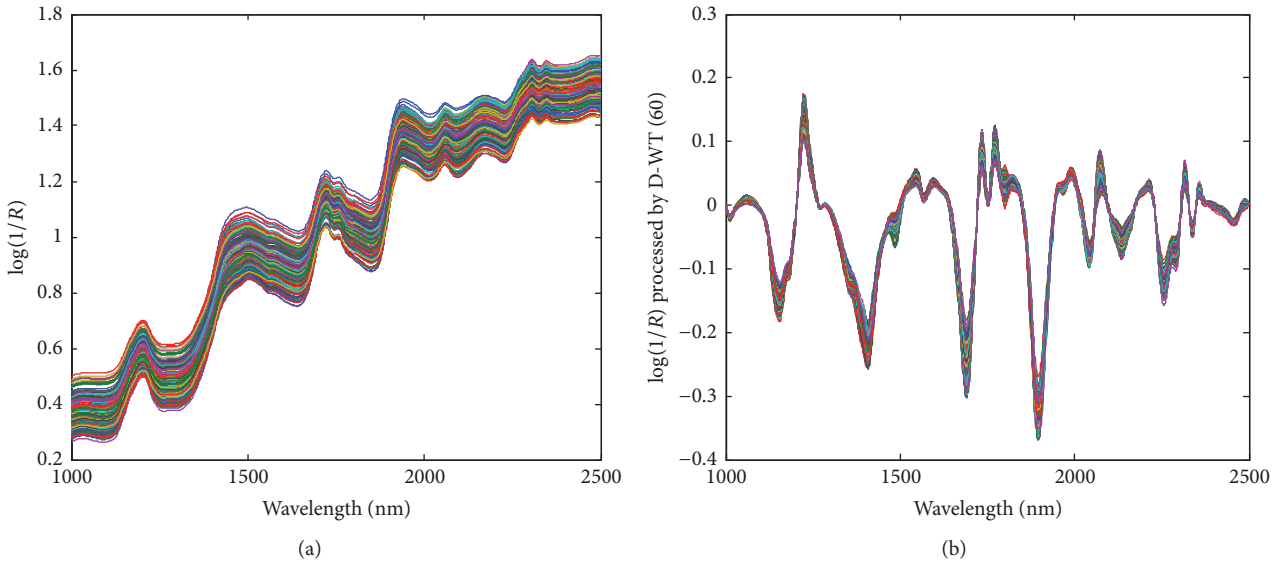

Figure 2: log(1/R) spectra from (a) original data and (b) processed data by the D-WT method.

samples) was used as the original sample set for determining the pretreatment method and screening for key variables. The prediction set (44 samples) was employed only for assessing the final performance of the models. The statistical results of the calibration and prediction sets are shown in Table 1. The protein, lipid, and moisture values in the calibration and prediction sets covered a sufficiently large range. More importantly, the range was greater in the calibration set than in the prediction set. These features are beneficial for developing a stable and robust model.

3.2. Determination of Spectral Preprocessing Method. The NIR spectral characteristics of the Fuzhu samples (Figure 2(a)) mainly represent the functional groups related to the content of moisture, proteins, lipids, and carbohydrates in the samples. These features commonly appear in the NIR spectra of soy foods acquired by diffuse reflectance techniques. For instance, they have been reported in the spectra of soybean [33], soya bean meal [34], and soybean flour [35]. The vibrational absorptions in NIR correspond to the vibrational transitions between the fundamental and higher-order energy levels and/or combination bands.

The original log(1/R) spectra of intact Fuzhu samples encode information on the effective path length and present the consistent offsets and biases in the baseline (Figure 2(a)). In this study, the characteristic signals were improved by various processes (MSC, SNV, derivative, SG smoothing, NDF, and D-WT). The RMSECV value was computed over the whole spectral range (1000–2499 nm) by the MCCV method, setting the number of MC simulations to 500 and the ratio of training samples (relative to the total number of samples) to 0.75. The MSC and SNV yielded similar RMSECV values (columns 3 and 4 in Table 2), indicating that both methods reduce the particle size effects; moreover, these two alternatives are interconvertible [8]. Judging from the results, 1D type is superior to 2D type under the current conditions. Derivative (1D or 2D) combined with smoothing methods (SG or NDF) slightly improved the precision of the models. D-WT (60) achieved the lowest RMSECV values among the tested preprocessing methods. Therefore, the D-WT method was selected for preprocessing the Fuzhu spectra and for variable selection and model development in further analysis. After processing by D-WT (Figure 2(b)), the spectral characteristic signals were highlighted in the wavelength ranges 1025–1050 nm, 1470–1590 nm, 1700–1800 nm, 1980–2140 nm, and 2290–2360 nm. Clearly, suppressing the noise corrects the drifting baseline and resolves the overlapping peaks.

3.3. Screening for Key Wavelengths

3.3.1. Variable Selection for Proteins. Figure 3(a) shows the average RI of each variable in the protein dataset in the wavelength range 1000–2499 nm, obtained by the MC-UVE method. The average RI (ordinate value) of each variable was computed from the absolute RIs (obtained by (1)) obtained in 100 iterations of MC-UVE. The dotted line indicates

Table 2: Lowest RMSECV values of quality parameters obtained by different pretreatment methods.

Quality parameters	Pretreatment method							
	Raw	MSC	SNV	1D	2D	1D-SG (7, 3)	2D-NDF (5, 5)	D-WT (60)
Protein	0.801	0.843	0.839	0.824	1.630	0.823	0.826	0.775
Lipid	0.728	0.687	0.687	0.833	1.390	0.796	0.793	0.681
Moisture	0.151	0.154	0.154	0.148	0.409	0.149	0.150	0.148

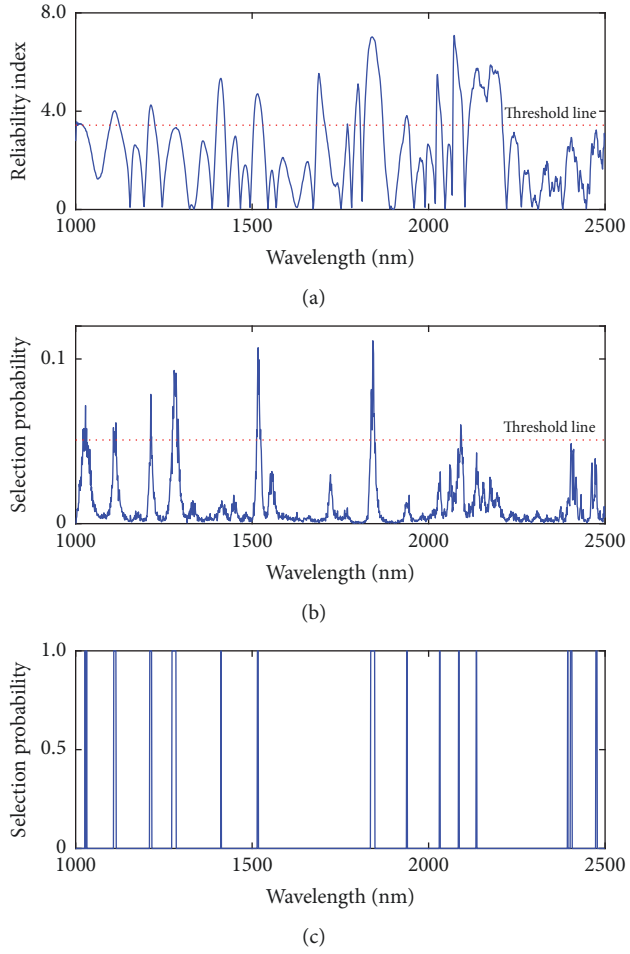

Figure 3: Results of variable screening of protein dataset by (a) MC-UVE, (b) RF, and (c) CARS.

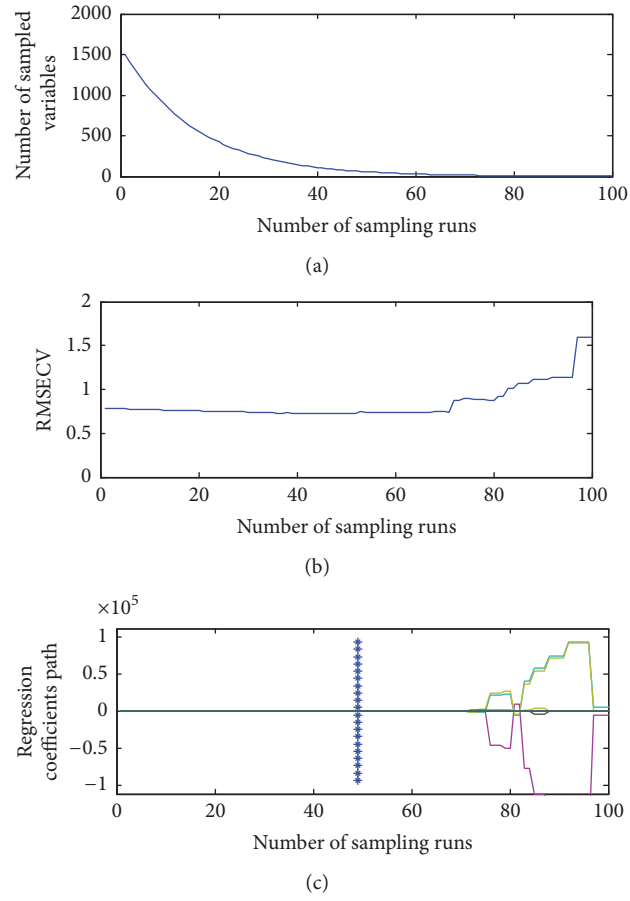

Figure 4: Changing trends in (a) number of sampled variables, (b) 5-fold RMSECV values, and (c) RC of each variable as the number of sampling runs increases. In (c), the line (marked by asterisks) indicates the optimal point where the 5-fold RMSECV values are minimized.

the threshold value, determined by the lowest RMSECV value as reported in the literature [28]. All RMSECV values were computed by the MCCV method using the parameters presented in the flow chart (Figure 1). Variables with RIs below the dotted line (RI = 3.42) were eliminated, and those with RIs above the dotted line were reserved for the PLS calculation. A total of 357 variables were selected by MC-UVE. The accepted variables were located around 1027, 1114, 1213, 1279, 1412, 1517, 1688, 1798, 1842, 1938, 2032, 2092, and 2113–2208 nm.

In the RF screening, the importance criterion of the variables was the mean SP value of 100 RF runs. Most of the variables scored below the mean SP line (see Figure 3(b)), but a small number of variables exhibited SP values above the line. The threshold value (dotted line in Figure 3(b)) was determined as mentioned for MC-UVE. Variables with SP values above the threshold (0.05) were regarded as the informative variables. This analysis identified 33 key variables for protein determination, concentrated in the spectral regions near 1021–1023 nm, 1025–1028 nm, 1031 nm, 1107 nm, 1110 nm, 1114 nm, 1212-1213 nm, 1215 nm, 1272–1286 nm, 1290-1291 nm, 1513–1524 nm, 1837–1847 nm, and 2091–2093 nm.

Figures 3(c) and 4 show the variable selection results of the protein dataset computed by CARS. The number of sampled variables changes with increasing number of sampling runs (Figure 4(a)). The decrease is rapid in the fast selection

stage of EDF and very slow in the refined selection stage [30]. In sampling runs 1–49, the 5-fold RMSECV values descend slightly as the uninformative variables are eliminated. In later sampling runs (50–72), they enter a relatively stable phase with no obvious changes and then rapidly increase in sampling runs 73–100 as useful information is lost (Figure 4(b)). Each line in Figure 4(c) records the RC of each variable in different sampling runs. The CARS analysis selected 61 variables for protein modeling, located in the spectral regions 1027 nm, 1031 nm, 1108–1114 nm, 1210–1215 nm, 1273–1284 nm, 1412 nm, 1515–1517 nm, 1836–1847 nm, 1938-1939 nm, 2031-2032 nm, 2090–2092 nm, 2135-2136 nm, 2394 nm, 2402–2406 nm, and 2474–2477 nm (Figure 3(c)).

Seven spectral regions around 1027, 1114, 1213, 1279, 1517, 1842, and 2092 nm (Figure 3) were selected by all three methods (MC-UVE, RF, and CARS). The spectral regions near 1412, 1938, and 2032 nm were chosen by both MC-UVE and CARS. The 2402–2406 nm and 2474–2477 nm regions were selected only by CARS, but their RIs in MC-UVE and SP values in RF were relatively high nonetheless. The lowest RMSECV values obtained by MC-UVE, RF, and CARS were 0.753, 0.741, and 0.723, respectively, less than that of PLS without variable selection (0.775; see Table 2). The numbers of selected variables observably decreased from 1500 to 357 in MC-UVE, to 55 in RF, and to 61 in CARS. This indicates that the variable selection methods effectively enhance the model performance and simplify the model dimensionality.

3.3.2. Variable Selection for Lipids. Figure 5(a) shows the mean RI of each variable in the lipid dataset in the wavelength range 1000–2499 nm, obtained by running MC-UVE 100 times. Variables with RIs above 6.5 were considered as informative variables; other variables were eliminated from the lipid data. A total of 134 variables concentrated in five main spectral regions (1189–1201 nm, 1417–1429 nm, 1581–1651 nm, 1749–1775 nm, and 1849–1858 nm) were selected for lipid modeling.

Figure 5(b) shows the average SP of each variable in the lipid dataset in the 1000–2499 nm range, obtained by running the FR 100 times. Variables with SP values below the dotted line (the threshold 0.035) were viewed as uninformative and eliminated from the lipid data; the remaining variables were reserved for PLS calibration. Finally, the RF analysis selected 58 variables in the spectral regions 1094–1098 nm, 1195–1206 nm, 1281–1292 nm, 1375–1385 nm, 1520–1528 nm, 1577–1594 nm, 1649 nm, and 1857–1860 nm (see Figure 5(b)).

Figure 5(c) shows the mean SP of each variable in the lipid dataset in the wavelength range 1000–2499 nm, computed after 100 sampling runs by the CARS method. The lowest 5-fold RMSECVs were obtained from 59 sampling runs. In total, CARS selected 31 variables for lipid modeling, located in the spectral regions 1094–1097 nm, 1201–1207 nm, 1383–1385 nm, 1421-1422 nm, 1483-1484 nm, 1519–1524 nm, 1649–1651 nm, 1858-1859 nm, 2017 nm, and 2061 nm (Figure 5(c)).

Three spectral regions around 1195, 1649, and 1858 nm (Figure 5) were selected by MC-UVE, RF, and CARS. Spectral regions around the ranges 1094–1097 nm, 1201–1207 nm, 1383–1385 nm, 1520–1524 nm, 1649 nm, and 1858-1859 nm,

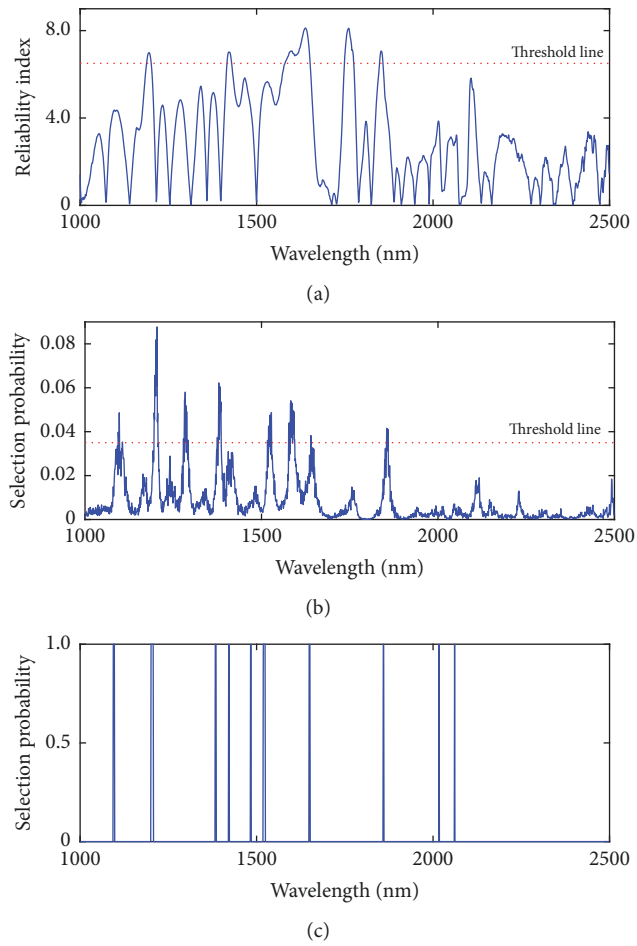

Figure 5: Results of variable screening of lipid dataset by (a) MC-UVE, (b) RF, and (c) CARS.

with relatively high RI, were selected by both RF and CARS. These algorithms work by different principles but yield similar results. The RMSECV values of MC-UVE (0.650), RF (0.636), and CARS (0.628) were all lower than the RMSECV value of PLS without variable selection (0.681; see Table 2). Moreover, the variables in the model calibration were drastically decreased from 1500 to 134 in MC-UVE, to 58 in RF, and to 31 in CARS.

3.3.3. Variable Selection for Moisture. Figure 6(a) shows the mean RI of each variable in the moisture dataset in the wavelength range 1000–2499 nm, obtained after 100 iterations of MC-UVE. The RI, RMSECV, and threshold values were calculated by the methods used for proteins and lipids. Variables with RI values below the threshold (RI = 5.5 in this case) were considered as uninformative and eliminated from the moisture data, leaving 112 variables for the PLS model development. These variables were concentrated in five spectral regions (1222–1239 nm, 1325–1349 nm, 1644–1674 nm, 1951–1966 nm, and 2047–2068 nm; see Figure 6(a)).

Figure 6(b) shows the average SP of each variable in the moisture dataset in the 1000–2499 nm range, obtained from 100 runs of FR. The dotted line indicates the threshold of

Table 3: Calibration and prediction results of protein, lipid, and moisture contents in Fuzhu obtained by different PLS models.

Quality parameter	Model	LVs	Variable number	R_C	RMSEC	R_P	RMSEP	RPD
Protein	MC-UVE-PLS	9	357	0.955	0.698	0.953	0.692	3.32
	RF-PLS	9	55	0.953	0.713	0.957	0.664	3.46
	CARS-PLS	9	61	0.958	0.677	0.958	0.656	3.51
Lipid	MC-UVE-PLS	8	134	0.945	0.607	0.962	0.475	3.62
	RF-PLS	10	58	0.948	0.585	0.966	0.442	3.89
	CARS-PLS	10	31	0.950	0.578	0.964	0.466	3.69
Moisture	MC-UVE-PLS	10	112	0.973	0.141	0.975	0.126	4.44
	RF-PLS	7	40	0.975	0.134	0.975	0.131	4.27
	CARS-PLS	5	29	0.974	0.137	0.976	0.123	4.55

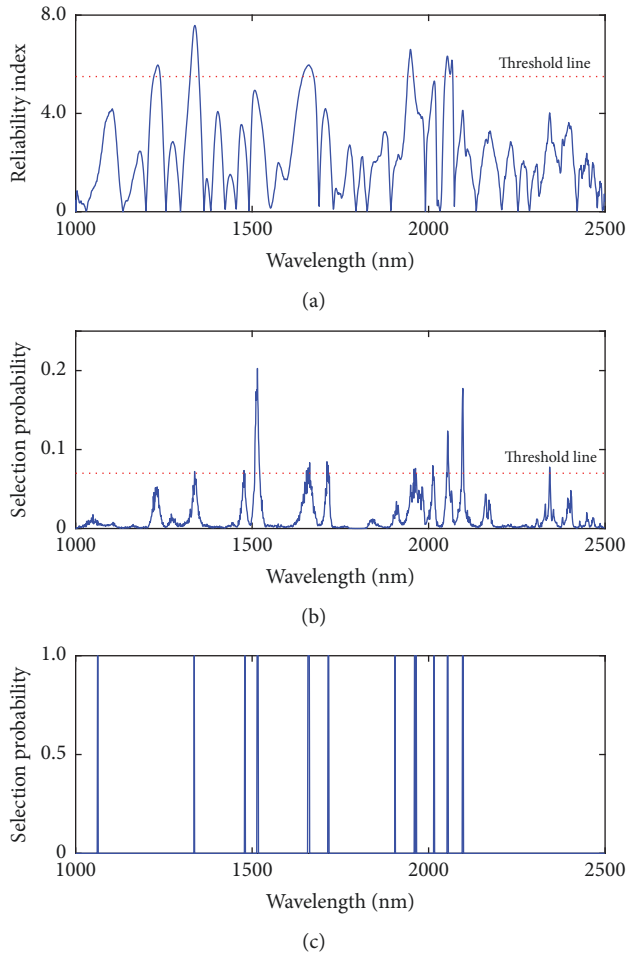

Figure 6: Results of variable screening of moisture dataset by (a) MC-UVE, (b) RF, and (c) CARS.

0.07, determined by the lowest RMSECV value. Variables with SP values below 0.07 were removed from the moisture data, whereas those with SP values above 0.07 were reserved for the PLS modeling. The RF analysis selected 40 key variables for moisture modeling, in the vicinities of 1336, 1479, 1515, 1662, 1716, 1964, 2012, 2054, 2097, and 2343 nm.

Figure 6(c) shows the average SP value of each variable in the moisture dataset in the wavelength range 1000–2499 nm, obtained from 100 sampling runs of CARS. The lowest 5-fold RMSECV values were obtained after 60 sampling runs. In total, CARS selected 29 variables for moisture evaluation, located in the spectral regions 1062-1063 nm, 1336 nm, 1479-1480 nm, 1514–1517 nm, 1658–1662 nm, 1715–1717 nm, 1905 nm, 1960-1961 nm, 1964 nm, 2015-2016 nm, 2053–2055 nm, and 2096–2098 nm.

Four spectral regions around 1336, 1662, 1964, and 2054 nm (Figure 6) were selected by all three methods (MC-UVE, RF, and CARS). Another four spectral regions (near 1479, 1515, 1716, and 2097 nm) were selected by both RF and CARS but rejected by MC-UVE because their RI values were below 5.5. Also noteworthy is the selection of the 1222–1239 nm region by MC-UVE alone, as well as the selection of the 1062-1063 nm region by CARS alone. The RMSECV values of MC-UVE (0.145), RF (0.140), and CARS (0.141) were only slightly lower than that of PLS over the whole spectral range (0.148; see Table 2). However, the number of variables was largely reduced from 1500 to 112 in MC-UVE, to 40 in RF, and to 29 in CARS.

3.4. Model Comparison and Assessment. The predictive performance of the models was evaluated on the prediction set containing 44 samples. Table 3 presents the results of the MU-UVE-PLS, RF-PLS, and CARS-PLS modeling for protein, lipid, and moisture of Fuzhu. The strong performance of the developed models is attributed to the high RPD values (3.32–3.51 for protein, 3.62–3.89 for lipid, and 4.27–4.55 for moisture). The developed model with a larger value of RPD (above 3.0) indicated the excellent ability of the model to precisely predict chemical compositions in new samples [36–38]. Comparing the model results of the quality parameters of intact Fuzhu, the parameters were most precisely estimated by the CARS-PLS, RF-PLS, and CARS-PLS models, with RPDs of 3.51, 3.89, and 4.55, for protein, lipid, and moisture, respectively.

Table 3 confirms that the variable selection methods not only provide satisfactory prediction accuracy, but also largely reduce the numbers of variables. This indicates that the MC-UVE, RF, and CARS methods both enhance the model performance and simplify the model complexity. Consequently, the best models were adopted in simultaneous determinations of the protein, lipid, and moisture contents in intact Fuzhu. Figure 7 shows the best calibration and prediction

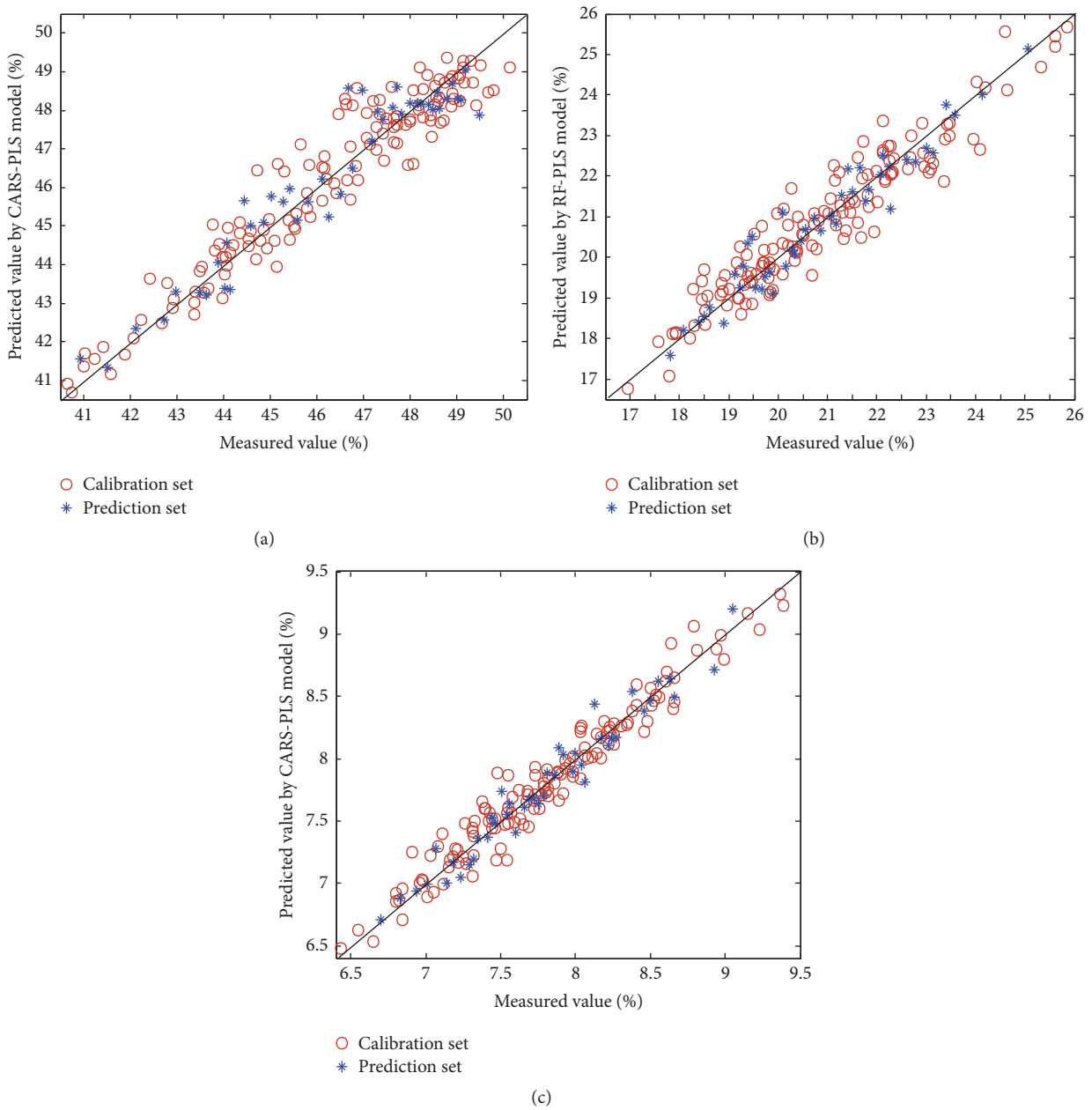

Figure 7: Scatter plots of predicted versus measured (a) protein, (b) lipid, and (c) moisture contents of Fuzhu.

results of proteins, lipids, and moisture obtained by the CARS-PLS, RF-PLS, and CARS-PLS models, respectively.

4. Conclusion

The protein, lipid, and moisture contents of intact Fuzhu were determined simultaneously by NIR spectroscopy in diffused reflectance mode. The efficiencies of various preprocessing methods were assessed by the RMSECV value computed by MCCV. Among these methods, D-WT achieved the optimal pretreatment. Three variable selection methods, MC-UVE, RF, and CARS, were then compared. Appropriate variable selection enhances the performance of a model. The protein, lipid, and moisture contents of Fuzhu were most precisely predicted by the CARS-PLS, RF-PLS, and CARS-PLS models, respectively. Combined with chemometrics, the NIR technique is suitable for quality control/evaluation of the traditional Chinese food "Fuzhu."

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant no. 31401579], China Scholarship Council (CSC), and Programs for Science and Technology Development of Henan Province of China [Grant no. 122102210247].

Supplementary Materials

The following are supplementary data to this article: (A) predicted and measured values of the best CARS-PLS models for protein; (B) predicted and measured values of the best RF-PLS models for lipid; (C) predicted and measured values of the best CARS-PLS models for moisture. *(Supplementary Materials)*

References

[1] L. C. Wu and R. P. Bates, "Soy protein-lipid films. 1. Studies on the film formation phenomenon," *Journal of Food Science*, vol. 37, no. 1, pp. 36–39, 1972.

[2] Y. Chen, S. Yamaguchi, and T. Ono, "Mechanism of the chemical composition changes of yuba prepared by a laboratory processing method," *Journal of Agricultural and Food Chemistry*, vol. 57, no. 9, pp. 3831–3836, 2009.

[3] C. A. Teixeira Dos Santos, M. Lopo, R. N. M. J. Páscoa, and J. A. Lopes, "A review on the applications of portable near-infrared spectrometers in the agro-food industry," *Applied Spectroscopy*, vol. 67, no. 11, pp. 1215–1233, 2013.

[4] J. U. Porep, D. R. Kammerer, and R. Carle, "On-line application of near infrared (NIR) spectroscopy in food production," *Trends in Food Science & Technology*, vol. 46, no. 2, pp. 211–230, 2015.

[5] L. Wang, D.-W. Sun, H. Pu, and J.-H. Cheng, "Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments," *Critical Reviews in Food Science and Nutrition*, vol. 57, no. 7, pp. 1524–1538, 2017.

[6] W.-H. Su, H.-J. He, and D.-W. Sun, "Non-Destructive and rapid evaluation of staple foods quality by using spectroscopic techniques: A review," *Critical Reviews in Food Science and Nutrition*, vol. 57, no. 5, pp. 1039–1051, 2017.

[7] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[8] M. Dhanoa, S. Lister, R. Sanderson, and R. Barnes, "The Link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR Spectra," *Journal of Near Infrared Spectroscopy*, vol. 2, no. 1, pp. 43–47, 2017.

[9] X. Shao and C. Ma, "A general approach to derivative calculation using wavelet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 69, no. 1-2, pp. 157–165, 2003.

[10] L. Nie, S. Wu, X. Lin, L. Zheng, and L. Rui, "Approximate derivative calculated by using continuous wavelet transform," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 2, pp. 274–283, 2002.

[11] C. B. Singh, R. Choudhary, D. S. Jayas, and J. Paliwal, "Wavelet analysis of signals in agriculture and food quality inspection," *Food and Bioprocess Technology*, vol. 3, no. 1, pp. 2–12, 2010.

[12] H. Xu, B. Qi, T. Sun, X. Fu, and Y. Ying, "Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears," *Journal of Food Engineering*, vol. 109, no. 1, pp. 142–147, 2012.

[13] Z. Huang, S. Sha, Z. Rong et al., "Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed," *Industrial Crops and Products*, vol. 43, no. 1, pp. 654–660, 2013.

[14] X.-D. Sun, M.-X. Zhou, and Y.-Z. Sun, "Variables selection for quantitative determination of cotton content in textile blends by near infrared spectroscopy," *Infrared Physics & Technology*, vol. 77, pp. 65–72, 2016.

[15] H.-L. Ma, J.-W. Wang, Y.-J. Chen, J.-L. Cheng, and Z.-T. Lai, "Rapid authentication of starch adulterations in ultrafine granular powder of Shanyao by near-infrared spectroscopy coupled with chemometric methods," *Food Chemistry*, vol. 215, pp. 108–115, 2017.

[16] C. M. Andersen and R. Bro, "Variable selection in regression-a tutorial," *Journal of Chemometrics*, vol. 24, no. 11-12, pp. 728–737, 2010.

[17] D. Jie, L. Xie, X. Fu, X. Rao, and Y. Ying, "Variable selection for partial least squares analysis of soluble solids content in watermelon using near-infrared diffuse transmission technique," *Journal of Food Engineering*, vol. 118, no. 4, pp. 387–392, 2013.

[18] B. A. J. G. Jacobs, B. E. Verlinden, E. Bobelyn et al., "Estimation of the prior storage period of lamb's lettuce based on visible/near infrared reflectance spectroscopy," *Postharvest Biology and Technology*, vol. 113, pp. 95–105, 2016.

[19] H.-D. Li, Q.-S. Xu, and Y.-Z. Liang, "Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification," *Analytica Chimica Acta*, vol. 740, pp. 20–26, 2012.

[20] X. Li, C. Sun, L. Luo, and Y. He, "Determination of tea polyphenols content by infrared spectroscopy coupled with iPLS and random frog techniques," *Computers and Electronics in Agriculture*, vol. 112, pp. 28–35, 2015.

[21] Y.-R. Zhao, K.-Q. Yu, X. Li, and Y. He, "Detection of fungus infection on petals of rapeseed (Brassica napus L.) using NIR hyperspectral imaging," *Scientific Reports*, vol. 6, Article ID 38878, 2016.

[22] C. Zhang, H. Ye, F. Liu, Y. He, W. Kong, and K. Sheng, "Determination and visualization of pH values in anaerobic digestion of water hyacinth and rice straw mixtures using hyperspectral imaging with wavelet transform denoising and variable selection," *Sensors*, vol. 16, no. 2, 2016.

[23] X. Wei, N. Xu, D. Wu, and Y. He, "Determination of Branched-Amino Acid Content in Fermented Cordyceps sinensis Mycelium by Using FT-NIR Spectroscopy Technique," *Food and Bioprocess Technology*, vol. 7, no. 1, pp. 184–190, 2014.

[24] X. Zhang, W. Li, B. Yin et al., "Improvement of near infrared spectroscopic (NIRS) analysis of caffeine in roasted arabica coffee by variable selection method of stability competitive adaptive reweighted sampling (SCARS)," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 114, pp. 350–356, 2013.

[25] Ministry of Health of the People's Republic of China, *Determination of protein in foods (GB 5009.5-2010)*, 2010.

[26] Ministry of Health of the People's Republic of China, *Determination of fat in foods (GB/T 5009.6-2003)*, 2003.

[27] Ministry of Health of the People's Republic of China (GB 5009.3-2010), 2010.

[28] W. Cai, Y. Li, and X. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 90, no. 2, pp. 188–194, 2008.

[29] Y.-H. Yun, H.-D. Li, L. R. E. Wood et al., "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 111, pp. 31–36, 2013.

[30] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Analytica Chimica Acta*, vol. 648, no. 1, pp. 77–84, 2009.

[31] P. C. Williams, "Implementation of near-infrared technology," in *In Near-infrared technology in the agricultural and food industries*, P. C. Williams and and K. H. Norris, Eds., pp. 145–171, AACC Inc., St. Paul, USA, 2001.

[32] J. Wang, J. Wang, Z. Chen, and D. Han, "Development of multi-cultivar models for predicting the soluble solid content and firmness of European pear (Pyrus communis L.) using portable vis–NIR spectroscopy," *Postharvest Biology and Technology*, vol. 129, pp. 143–151, 2017.

[33] D. S. Ferreira, O. F. Galão, J. A. L. Pallone, and R. J. Poppi, "Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples," *Food Control*, vol. 35, no. 1, pp. 227–232, 2014.

[34] S. A. Haughey, S. F. Graham, E. Cancouët, and C. T. Elliott, "The application of Near-Infrared Reflectance Spectroscopy (NIRS) to detect melamine adulteration of soya bean meal," *Food Chemistry*, vol. 136, no. 3-4, pp. 1557–1561, 2013.

[35] L. P. Brás, S. A. Bernardino, J. A. Lopes, and J. C. Menezes, "Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, no. 1, pp. 91–99, 2005.

[36] P. C. Williams and D. C. Sobering, "How do we do it: a brief summary of the methods we use in developing near infrared calibrations," in *In Near infrared spectroscopy: the future waves*, A. M. C. Davies and P. C. Williams, Eds., pp. 185–188, NIR Publications, Chichester, UK, 1996.

[37] W. Saeys, A. M. Mouazen, and H. Ramon, "Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy," *Biosystems Engineering*, vol. 91, no. 4, pp. 393–402, 2005.

[38] M. W. Davey, W. Saeys, E. Hof, H. Ramon, R. L. Swennen, and J. Keulemans, "Application of visible and near-infrared reflectance spectroscopy (vis/NIRS) to determine carotenoid contents in banana (musa spp.) fruit pulp," *Journal of Agricultural and Food Chemistry*, vol. 57, no. 5, pp. 1742–1751, 2009.